文章编号：1003-0530(2024)07-1341-13

基于DRL的定向网络时隙复用和功率控制协议

梁仕杰[1,2] 赵海涛*[2] 张 姣[2] 王海军[2] 魏急波[2] 王俊芳[1]
（1. 中国电子科技集团公司第五十四研究所，河北石家庄 050081；
2. 国防科技大学电子科学学院，湖南长沙 410073）

摘 要：近年来，无人机网络逐渐地广泛应用于各行各业，对无人机网络能提供的网络容量提出了更高的要求。定向天线结合无人机网络构成定向无人机网络以增加网络资源应对无人机网络中各个节点对网络有限通信资源的竞争造成网络容量低的问题。定向无人机网络通过定向天线的空间复用能力可以提高网络的时隙利用效率。针对TDMA协议在定向组网中时隙利用率过低导致网络容量受限的问题，该文提出了一种基于深度Q网络（DQN）的定向无人机网络时隙复用和功率控制协议。为了提高时隙利用率，考虑在单位时隙进行多个链路通信以实现时隙资源的复用。然而多个链路在同一个时隙通信会产生链路间的干扰，如何在考虑链路间相互干扰的情况下控制功率提高网络的容量是时隙复用研究的重点问题。为了解决该问题，首先考虑以功率要求和每条链路最小信道容量为约束，考虑相较于其他研究更为复杂更符合实际的链路互干扰模型，建模问题为最大全网容量问题。然后为了构建链路间的更复杂的互干扰环境，将多个链路的瞬时信道信息、定向增益状态融入到DQN框架的状态中，DQN的奖励为高于最小信道容量的链路信道容量的和。最后，将每个时隙的优化问题扩展到每一帧的优化问题，并利用多个DQN进行求解。仿真结果表明，在保证每个被分配时隙的最小信道容量前提下，相较于对比方法网络容量有了很大的提升。

关键词：时分多址协议；定向无人机网络；深度Q网络；时隙复用；功率控制
中图分类号：TN914.52 **文献标识码：**A **DOI：**10.16798/j.issn.1003-0530.2024.07.015

引用格式：梁仕杰，赵海涛，张姣，等. 基于DRL的定向网络时隙复用和功率控制协议[J]. 信号处理，2024，40(7)：1341-1353. DOI：10.16798/j.issn.1003-0530.2024.07.015.
Reference format：LIANG Shijie，ZHAO Haitao，ZHANG Jiao，et al. Directional network slot reuse and power control protocol based on DRL[J]. Journal of Signal Processing，2024，40(7)：1341-1353. DOI: 10.16798/j. issn. 1003-0530.2024.07.015.

Directional Network Slot Reuse and Power Control Protocol Based on DRL

LIANG Shijie[1,2] ZHAO Haitao*[2] ZHANG Jiao[2] WANG Haijun[2] WEI Jibo[2] WANG Junfang[1]
（1. The 54th Research Institute of China Electronics Technology Group Corporation，Shijiazhuang，Hebei 050081，China；
2. College of Electronic Science and Technology，National University of Defense Technology，Changsha，Hunan 410073，China）

Abstract: In recent years，unmanned aerial vehicle（UAV）networks have been progressively and extensively employed in various industries，which places higher demands on the network capacity that drone networks can provide. Directional antennas combined with drone networks form directional drone networks to address the problem of low network capacity caused by the competition for limited communication resources among nodes in UAV networks. Directional UAV networks can improve slot utilization through the spatial reuse capability of directional antennas. In response to the problem of low slot utilization and limited network capacity within directional networks employing the TDMA protocol，this study pro-

收稿日期：2023-09-15；修回日期：2023-12-10
*通信作者：赵海涛 haitaozhao@nudt.edu.cn *Corresponding Author：ZHAO Haitao，haitaozhao@nudt.edu.cn
基金项目：国家自然科学基金重点项目(61931020)；国家自然科学基金(62001483)；湖南省自然科学基金杰青项目(2022JJ10068)
Foundation Items：The National Natural Science Foundation of China（61931020，62001483）；The Project supported by Provincial Natural Science Foundation of Hunan（2022JJ10068）

poses a protocol for slot reuse and power control in directional UAV networks based on deep Q-networks (DQNs). To enhance slot utilization, we consider multiple links communicating in a single slot to achieve slot reuse. However, the simultaneous communication of multiple links in a single slot introduces inter-link interference. Managing power control to increase network capacity while considering this interference among links is a key focus in slot reuse research. To address this problem, we first consider imposing constraints based on power requirements and the minimum channel capacity for each link. Considering a more complex and practical link interference model compared with other studies, the problem is formulated as a maximum overall network capacity problem. Subsequently, for a more intricate inter-link interference environment, the instantaneous channel information and directional gain states of multiple links are incorporated into the state of the DQN framework. The reward for the DQN is defined as the sum of the channel capacities of links that exceed the minimum channel capacity. Finally, by extending the optimization problem of each time slot to that of each frame, multiple DQNs are utilized. Simulation results demonstrate that, while ensuring the minimum channel capacity of each allocated slot, the proposed method significantly increases network capacity compared with the benchmark methods.

Key words: time division multiple access (TDMA) protocol; directional UAV network; deep Q-network (DQN); slot reuse; power control

1 引言

近年来,由于无人机(Unmanned Aerial Vehicle, UAV)具有灵活性、高机动性等特点,越来越多的用于执行监测、侦查、救援等任务[1-2]。多无人机通过组网以扩大执行任务的覆盖范围。然而,无人机网络有限的通信资源面临着网内节点的竞争和干扰,这对网络容量产生了很大的负面影响。将定向天线应用到无人机组网中构建定向无人机网络,可以有效提升网络容量[3-4]。

不同于全向通信对周围非目标节点的干扰增益相同,定向通信只会在通信的角度范围内产生高增益干扰。因此在定向无人机网络的内部干扰主要有两种:主瓣干扰和旁瓣干扰。以图1为例,节点C受到节点A高增益的主瓣干扰,节点B受到节点A较小的旁瓣干扰。定向天线在一个角度进行通信的特性使其受到主瓣干扰的概率很低,而旁瓣干扰增益很小,因此定向无人机网络中可以使多个链路在一个时隙工作从而提升网络容量。这种能力称为定向无人机网络的时隙复用能力。然而TDMA协议将每个时隙分配给一个节点或者链路,无法利用定向无人机网络的时隙复用能力提高信道容量。因此,在TDMA协议框架下进行多链路复用必然会引入网络内部干扰,如何合理的进行时隙复用和控制节点的发射功率是提升网络容量的关键。

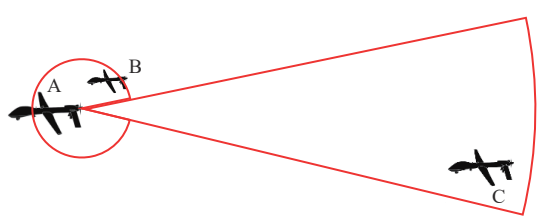


图1 主瓣干扰和旁瓣干扰

Fig. 1 Main and side lobe interferences

当前,利用定向天线的空间复用实现时隙复用已有一些很好的研究[5-9]。文献[5]提出了一种基于六边形图案的时隙的空间复用方案以支持大型无人机编队的传输。文献[6]提出了一种基于STDMA(Spatial-Time Division multiple Access)的调度算法,使用STDMA来提高系统吞吐量,允许非干扰和干扰链路同时传输以提高毫米波网络的资源利用效率,但是其考虑的定向天线模型和干扰模型较简单。文献[7]提出了基于混合整数线性规划问题(Mixed Integer Linear Programming, MILP)的多时隙调度方案,用于提高毫米波无线个人局域网(Wireless Personal Area Networks, WPAN)的能量调度效率和延迟公平性。同时,该文章提出利用强化学习多时隙的调度框架,用强化学习计算进行最佳速率的请求。文献[8]提出了将利用强化学习的方式实现定向链路并行运行,但是并未考虑对功率的调控。文献[9]将无人机自组网的时隙、方向和功率分配问题建模为一个混合整数非线性规划问题(Mixed Integer Nonlinear Programming Problem, MINLP)问题,并提出了基于对偶的迭代搜索算法(Dual-Based Iterative Search Algorithm, DISA)和序贯穷举资源分配算法(Sequential Exhausted Allocation Algorithm, SEAA)来解决所提出的混合整数非线性规划问题,但并未实现多定向链路在同一时隙的复用。立足现有的研究内容,我们通过进一步研究以实现在复杂干扰情况下定向网络时隙复用和功率控制的二维资源调度。

当前一些针对功率优化问题的研究利用传统

方法进行求解[9-15]，例如采用拉格朗日松弛法、拉格朗日对偶分解、连续凸逼近法、贪心算法等。除了文献[9]，其他均是面向全向网络的功率控制研究，并且这些传统的方法无法适应动态的场景。而功率控制会影响互干扰环境，加之信道状态是动态的，因此导致动态的环境。针对动态环境的问题，强化学习算法能够通过与环境的交互自主学习，并在没有明确的监督信号的情况下进行决策。强化学习通过探索策略，可以主动探索未知领域并根据环境进行新的动作。相较于传统方法，这种探索能力使得强化学习在面对复杂的环境和多样的任务时具备了优势。然而强化学习难以处理高维状态空间的问题，深度强化学习（Deep Reinforcement Learning，DRL）能够利用深度神经网络作为值函数估计器可以弥补强化学习在处理高维状态的复杂环境时的缺陷[16]。因此，很多研究利用深度强化学习的框架以解决复杂环境下的功率分配问题[17-22]。文献[17]提出了一种新的分布式分层深度强化学习（DHDRL）以实现功率分配。文献[18]提出了一种回声状态网络的深度强化学习算法进行功率控制减少了无人机对地面的干扰。文献[19]利用深度Q网络（Deep Q-Network，DQN）求解关于控制功率的优化问题。文献[20]提出了两种互补的深度强化学习算法。这两种算法同时训练和执行以实现子带和功率的分配。文献[21]利用深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）实现控制功率从而提高无人机自组网回程传输速率。文献[22]提出了一种新的深度强化学习方法——软演员-评论家算法（Soft Actor-Critic，SAC）来进行功率控制从而实现最大化吞吐量。

上述基于强化学习的功率控制研究均是在全向网络下实现功率控制，并且没有考虑最小信道容量的约束。通过设置最小信道容量可以保证实际应用场景中的业务速率需求。

针对当前TDMA协议在定向无人机网络时隙利用率过低的问题，我们提出了基于DQN的时隙复用和功率控制协议，在保证每个链路的最小信道容量的前提下，实现了定向网络时隙的复用和功率控制提升了全网的信道容量。为了贴近更实际的场景，我们考虑了主瓣-旁瓣干扰，旁瓣-主瓣干扰，主瓣-主瓣干扰和旁瓣-旁瓣干扰等四种干扰和动态的信道环境。首先利用TDMA协议[23]为每个链路分配一个时隙。然后，在每个时隙构建优化问题：以功率和最小信道容量为约束实现全网容量的最大化目标。在定向无人机网络内部复杂的干扰环境下，将复杂的干扰环境加入到DQN框架的状态中，以最大化全网容量的优化目标和最小功率限制联合设计的DQN奖励实现了多个链路在一个时隙的复用和功率控制。大量仿真实验验证了本文所提方法的有效性和先进性。最后，仿真实验表明，本文所提方法相比于基准算法可以获得更高的网络容量。

2　系统与问题建模

2.1　系统模型

考虑在某区域内部署多个无人机节点和一个地面控制站构建定向无人机集群网络，如图2所示。其中，每个无人机节点均配置两套工作在不同的频段的定向天线和收发终端，分别负责无人机间和无人机与地面站之间的数据通信。地面站作为中心控制节点对定向无人机网络进行集中控制、任务调度和资源分配。无人机与地面站之间的通信采用OFDM波形，其上下行链路协议参照OFDMA协议[24]。在本文中，考虑定向天线水平方向窄波束夹角，俯仰方向的宽波束夹角，并认为只要在水平方向覆盖节点即可实现俯仰方向的覆盖。因此本文只需要考虑水平方向的夹角覆盖产生的干扰。

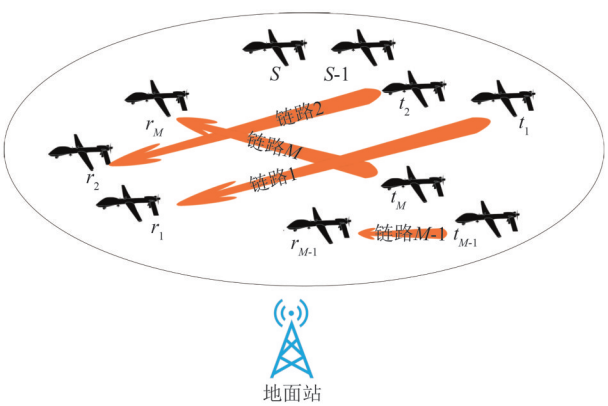


图2　定向无人机网络通信示意图

Fig. 2　Illustration of directional UAV network communication

第i条链路的发射节点和接收节点分别为t_i和r_i，节点t_i的定向天线发射模型如式（1）所示，其中φ_{t_i}和$\tilde{\varphi}_{t_i}$分别表示节点t_i主瓣指向方向和通信方向，θ表示定向天线主瓣角度，G_{t_i}单位为dB。

$$G_{t_i}(\varphi_{t_i},\tilde{\varphi}_{t_i})=\begin{cases}G_{\text{main}}, & \left|\varphi_{t_i}-\tilde{\varphi}_{t_i}\right|\leqslant\dfrac{\theta}{2}\\ 0, & \text{otherwise}\end{cases}\tag{1}$$

接收节点r_i的定向天线接收增益模型如式(2)所示，φ_{r_i}，$\tilde{\varphi}_{r_i}$分别表示接收端r_i的主瓣指向方向和通信方向。

$$G_{r_i}(\varphi_{r_i},\tilde{\varphi}_{r_i})=\begin{cases}G_{\text{main}}, & \left|\varphi_{r_i}-\tilde{\varphi}_{r_i}\right|\leqslant\dfrac{\theta}{2}\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

在通信阶段，如图3所示，M个链路的收发节点的主瓣均指向对方，即通信方向和主瓣指向方向相同。因此，$\varphi_{t_i}=\tilde{\varphi}_{t_i}$，$\varphi_{r_i}=\tilde{\varphi}_{r_i}$。$\tilde{\varphi}_{r_i}$和$\tilde{\varphi}_{t_i}$的关系如式(3)所示。

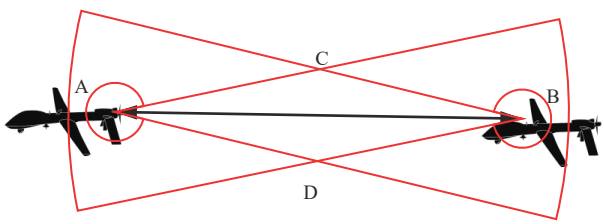


图3 定向天线波束指向示意图

Fig. 3 Illustration of directional antenna beam pointing

$$\tilde{\varphi}_{r_i}=\begin{cases}\tilde{\varphi}_{t_i}+180°, & \tilde{\varphi}_{t_i}+180°<360°\\ \tilde{\varphi}_{t_i}-180°, & \tilde{\varphi}_{t_i}+180°\geqslant 360°\end{cases} \tag{3}$$

每个链路的发射节点均被其他链路的接收节点视为干扰节点，t_j对r_i的干扰增益如式(4)所示。φ_{t_j}表示t_j的通信方向，φ_{t_j,r_i}表示t_j指向r_i的方向。

$$G_{t_j,r_i}(\varphi_{t_j},\varphi_{t_j,r_i})=\begin{cases}G_{\text{main}}, & \left|\varphi_{t_j}-\varphi_{t_j,r_i}\right|\leqslant\dfrac{\theta}{2}\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

每个链路的接收节点被其他链路的发射节点干扰的情况下，每个接收节点对干扰节点的接收干扰增益如式(5)所示。φ_{r_i,t_j}表示r_i指向t_j的方向。

$$G_{r_i,t_j}(\varphi_{r_i},\varphi_{r_i,t_j})=\begin{cases}G_{\text{main}}, & \left|\varphi_{r_i}-\varphi_{r_i,t_j}\right|\leqslant\dfrac{\theta}{2}\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

由式(4)和式(5)可以得知，干扰主要分为四种情况：

干扰节点和被干扰节点均在对方的主瓣覆盖范围内，这时干扰的定向增益为两个主瓣增益的和；

干扰节点在被干扰节点的主瓣范围内，而被干扰节点不在干扰节点的主瓣范围内，这时干扰的定向增益为干扰节点的旁瓣增益和被干扰节点主瓣增益的和；

干扰节点不在被干扰节点的主瓣范围内，而被干扰节点在干扰节点的主瓣范围内，这时干扰的定向增益为干扰节点的主瓣增益和被干扰节点旁瓣增益的和；

干扰节点和被干扰节点均不在对方的主瓣覆盖范围内，这时干扰的定向增益为两个旁瓣增益的和。

结合式(1)(2)(4)和(5)，假设每个时隙的瞬时信道状态信息已知[25]，得到考虑干扰的情况下在时隙k链路i接收端的信干噪比(SINR)，如式(6)所示。其中，h_i^k表示在时隙k中t_i到r_i的瞬时信道状态信息(Channel State Information，CSI)，$h_{j,i}^k$表示在时隙k中干扰节点t_j到r_i的瞬时信道状态信息。p_i^k和p_j^k分别表示在时隙k中t_i的发射功率和t_j的发射功率。$\boldsymbol{M}=\{1,2,3,\cdots,M\}$表示所有链路的集合。$N_0$表示加性高斯白噪声。

$$\rho_i^k=\frac{h_i^k p_i^k 10^{\frac{G_{t_i}}{10}}10^{\frac{G_{r_i}}{10}}}{N_0+\sum_{j\neq i}^{j\in M}h_{j,i}^k p_j^k 10^{\frac{G_{t_j,r_i}}{10}}10^{\frac{G_{r_i,t_j}}{10}}} \tag{6}$$

根据式(6)可以得到链路i在时隙k的归一化信道容量：

$$C_i^k=\log_2(1+\rho_i^k) \tag{7}$$

定向无人机网络和全向无人机网络不同，定向无人机网络的定向增益使得无人机间的通信可以被视为一个具有一定角度的通信链路，如图2所示。在主瓣范围外发射节点对其他节点干扰很小，可以将关注点集中在链路上。因此本文对定向网络中链路占据的时隙和发射功率进行优化。

2.2 问题建模

在图2中，共有S个无人机中的$2M(2M\leqslant S)$个无人机构成M个通信链路准备进行通信。我们设定一帧有M个时隙。为了实现一帧内M个时隙的网络容量的最大化和公平的时隙分配，首先采用TDMA协议每个链路占据一个时隙[23]，以保证每个链路在一帧内通信一次。然后保证每个链路在被分配时隙的信道容量大于最小信道容量的前提下，利用定向天线的时隙复用能力通过功率控制实现网络容量的最大化。根据TDMA协议，假定时隙k分配给链路n，定义的如式(8)所示的优化问题。通过求解时隙k的发射功率$\boldsymbol{P}^k$以实现时隙k的最大网络容量$\sum_{i,n\in\boldsymbol{M}}C_i^k$。$C1$对未分配该时隙的链路的功率进行限制，功率为0表示在该时隙不工作。$C2$表示，只要该链路在时隙k工作该链路的信道容量就必须大于最小信道容量$C_{\min}$以保证业务速率要求。$C3$表示被分配该时隙的链路n必须处于工作状态，$C4$保证了链路n信道容量的最小值。$C_{\min}$设定与物理层和业务需求相关，针对不同的场景按需设置。

$$\max_{\boldsymbol{P}^k} \sum\nolimits_{i,n\in\boldsymbol{M}} C_i^k$$
$$\text{s.t. } C1:\ 0\leqslant p_i^k|_{i\neq n}\leqslant P_{\max}$$
$$C2:\ C_i^k|_{i\neq n}\geqslant C_{\min},\ \text{if } p_i^k\neq 0$$
$$C3:\ 0<p_n^k\leqslant P_{\max}$$
$$C4:\ C_n^k\geqslant C_{\min} \tag{8}$$

根据式(7)，在M个时隙内全网所有链路的信道传输总容量如式(9)所示，其中$\boldsymbol{T}=\{1,2,3,\cdots,M\}$表示所有时隙的集合。

$$C^M=\sum\nolimits_{k\in\boldsymbol{T}}\sum\nolimits_{i\in\boldsymbol{M}} C_i^k \tag{9}$$

为了实现式(9)最大化，在每个经过TDMA协议分配后的时隙进行优化问题(8)求解。式(8)的问题本质上是一个非凸非线性的规划问题，传统方法中一般采用将其转化为凸问题后进行求解[26]。而随着智能算法的发展，可以利用智能算法对该问题进行近似求解。同时由于信道环境动态变化且状态信息复杂，因此本文提出了基于DQN求解问题(8)的时隙复用和功率控制协议。

3　基于DQN的时隙复用和功率控制协议

3.1　深度Q网络

如图4所示，强化学习的框架由智能体和环境相互作用组成。在时间t，执行强化学习算法的智能体通过对环境进行观察获得状态$\boldsymbol{s}^t\in\boldsymbol{S}$，通过执行动作$\boldsymbol{a}^t\in\boldsymbol{A}$和环境进行交互获得时隙$t$的奖励$r^t$和下一个状态$\boldsymbol{s}^{t+1}\in\boldsymbol{S}$。$\boldsymbol{S}$和$\boldsymbol{A}$分别表示是所有状态和动作的集合。智能体采取与环境的持续交互的方式，根据来自环境的反馈调整策略。为了使用强化学习进行求解，需要优化的问题通常被表述为如图5所示的马尔可夫决策过程(Markov Decision Process，MDP)。在MDP中，t+1时刻的状态$\boldsymbol{S}^{t+1}$只与t时刻的状态$\boldsymbol{S}^t$和动作$\boldsymbol{a}^t$相关。

Q-learning算法是当前最流行的用来处理马尔可夫决策问题(MDP)强化学习算法之一[27]。当状态空间连续等原因造成状态空间无限大时，Q-learning建立的Q值表变得非常庞大。为解决Q-learning这一缺陷，DQN将Q-learning和深度神经网络结合利用深度神经网络作为值函数近似器的输出来代替查找Q值表所得到的Q值。根据优化问题设计DQN的奖励、状态和动作以实现通过DQN来求解优化问题。

DQN中的累计奖励如式(10)所示，其中γ是用来平衡未来和现在奖励的折扣因子，r表示奖励。

$$R^t=\sum\nolimits_{\tau}^{\infty}\gamma^{\tau}r^{t+\tau+1} \tag{10}$$

DQN利用如图6所示的深度Q网络来逼近Q函数。在确定策略π的条件下，智能体在状态s的条件下执行动作a的Q函数如式(11)所示，其中$\boldsymbol{\theta}$是迭代的神经网络的权重向量，$\mathbb{E}[\cdot]$表示进行期望计算。DQN目标是通过和未知的环境的交互最大化Q函数。

$$Q_\pi(s,a;\ \boldsymbol{\theta})=\mathbb{E}_\pi[R^t|\boldsymbol{s}^t=\boldsymbol{s},\boldsymbol{a}^t=\boldsymbol{a}] \tag{11}$$

DQN通过多次迭代获得近似最优策略为：

$$\pi^*(\boldsymbol{s},\boldsymbol{a})=\arg\max_{\boldsymbol{a}} Q_\pi(\boldsymbol{s},\boldsymbol{a}) \tag{12}$$

根据贝尔曼方程[28]，最大化式(11)可以描述为式(13)所示，其中y^t是最优的Q值。

$$y^t=r^t+\gamma\max_{\boldsymbol{a}} Q(\boldsymbol{s}^{t+1},\boldsymbol{a};\ \boldsymbol{\theta}^t) \tag{13}$$

在DQN中，最主要的任务是对神经网络进行训练，以使其输出可以近似于Q函数。在神经网络训练中，采用基于随机梯度下降的方法令当前的$Q(s^t,a^t;\ \boldsymbol{\theta}^t)$逐渐逼近$y^t$，最小化如式(14)所示的损

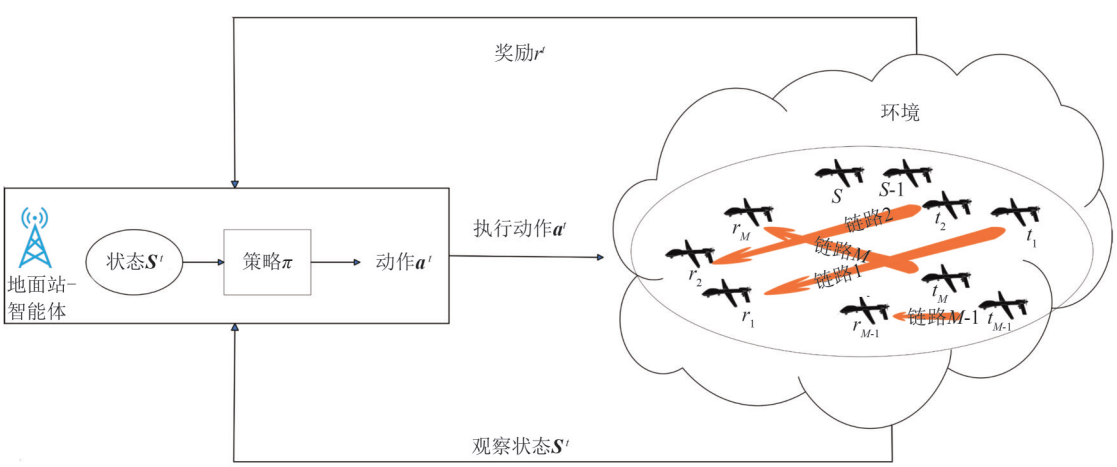


图4　定向无人机网络环境下的强化学习

Fig. 4　Reinforcement learning in the environment of the directional UAV network

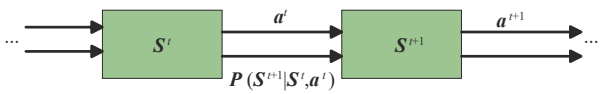


图5 马尔可夫决策过程

Fig. 5 Markov decision process

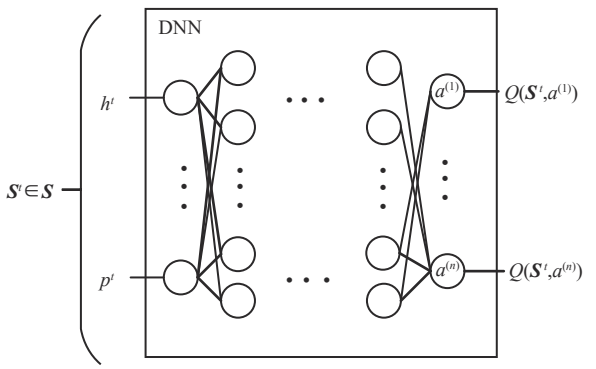


图6 深度Q网络结构

Fig. 6 Deep Q-network structure

失函数从而更新神经网络的参数$\boldsymbol{\theta}$。参数$\boldsymbol{\theta}$更新公式如式(15)所示,其中η为学习率。

$$L(\boldsymbol{\theta})=E[y^t-Q(\boldsymbol{s}^t,\boldsymbol{a}^t;\ \boldsymbol{\theta}^t)] \tag{14}$$

$$\boldsymbol{\theta}^{t+1}=\boldsymbol{\theta}^t+\eta(y^t-Q(s^t,a^t;\ \boldsymbol{\theta}^t))\nabla Q(s^t,a^t;\ \boldsymbol{\theta}^t) \tag{15}$$

神经网络的训练数据源于智能体与环境交互时所产生的经验组,这些经验被储存在经验缓冲池中,每个经验由当前状态、选择的动作、获得的奖励和下一个状态组成,即(s^t,a^t,R^t,s^{t+1})。DQN算法在训练神经网络时在经验缓冲池中进行随机采样训练以减少和环境的交互次数提高训练效率。

问题(8)的优化目标是最大化每个时隙的网络容量。为此将在分配给链路n的时隙t的奖励设计为式(16)。通过DQN最大化式(16)的含义是在时隙t保证链路n信道容量和功率不为0的链路信道容量大于最小信道容量的前提下,最大化全网功率不为0的大于最小信道容量链路的信道容量的和,即最大化全网有效信道容量。其中$\boldsymbol{J}$表示除链路n外功率大于零的链路集合,j是J中的元素。m是所有链路$\boldsymbol{M}$的元素。

$$r^t=\begin{cases}0, C_n^t<C_{\min}\ \text{or}\ \ 0<C_j^t<C_{\min}|_{\forall j\in \boldsymbol{J}}\\ \sum\limits_{m\in M}S_m^t, \qquad\qquad \text{otherwise}\end{cases} \tag{16}$$

$$S_m^t=\begin{cases}C_m^t, \text{if}\ C_m^t\geqslant C_{\min}|_{\forall m\in \boldsymbol{M}}\\ 0,\ \ \text{if}\ C_m^t<C_{\min}|_{\forall m\in \boldsymbol{M}}\end{cases} \tag{17}$$

由于只关心当前时隙的信道容量,因此本文将式(10)的折扣因子γ设为0,根据最大化式(11)的目标我们可以得到新的Q函数:

$$\max Q_\pi=\max_{\boldsymbol{a}\in \boldsymbol{A}}\mathbb{E}_\pi[r^t|\boldsymbol{s}^t=\boldsymbol{s},\boldsymbol{a}^t=\boldsymbol{a}] \tag{18}$$

根据式(16)和(17),我们可以得到时隙t奖励只与信道状态信息$\boldsymbol{H}^t=\{h_{i,j}^t\}_{i\in \boldsymbol{M},\,j\in \boldsymbol{M}}$,定向增益$\boldsymbol{G}^t=\{G_{t_i}^t,G_{r_i}^t,G_{t_j,r_i}^t,G_{r_i,t_j}^t\}_{i\in \boldsymbol{M},\,j\in \boldsymbol{M}}$和功率$\boldsymbol{P}^t=\{\boldsymbol{P}_i^t\}_{i\in \boldsymbol{M}}$有关,其中$h_{i,j}^t$表示在时隙$t$时$t_i$到$r_i$的信道状态信息,因此我们可以得到:

$$\max Q_\pi=\max_{0\leqslant \boldsymbol{P}^t\leqslant P_{\max}}\mathbb{E}_\pi[r^t|\boldsymbol{H}^t,\boldsymbol{G}^t,\boldsymbol{P}^t] \tag{19}$$

在DQN执行期间,策略是确定性的,因此式(19)可以写为:

$$\max Q=\max_{0\leqslant \boldsymbol{P}^t\leqslant P_{\max}}r^t(\boldsymbol{H}^t,\boldsymbol{G}^t,\boldsymbol{P}^t) \tag{20}$$

为了辅助DQN快速找到最优解,在状态信息中增加上一个时隙的奖励r^{t-1}和动作$\boldsymbol{P}^{t-1}$,因此式(20)可以改写为:

$$\max Q=\max_{0\leqslant \boldsymbol{P}^t\leqslant P_{\max}}r^t(\boldsymbol{H}^t,\boldsymbol{G}^t,\boldsymbol{P}^t,r^{t-1},\boldsymbol{P}^{t-1}) \tag{21}$$

由于折扣因子γ为0,因此可以对式(13)进行简化得到$y^t=r^t$,同时经验回放可以简化为$(\boldsymbol{s}^t,\boldsymbol{a}^t,r^t,\boldsymbol{s}^{t+1})$。

3.2 定向无人机网络DQN设计

在如图2所示的定向无人机网络中,链路之间的通信可以被看作是一个智能体,因此优化定向无人机网络资源的问题可以被视为一个多智能体优化问题。然而,无人机作为资源和功耗受限的平台,难以提供多智能体训练的计算资源和功耗需求。因此本文利用地面站进行集中式的训练,然后通过分发使网络中的每个无人机共享学习到的策略。DQN经验回放的状态、动作和奖励设计如下:

状态:由于信道的幅度可能以数量级变化,因此对信道的CSI信息$\boldsymbol{H}^t=\{h_{i,j}^t\}_{i\in \boldsymbol{M},\,j\in \boldsymbol{M}}$进行如式(22)的处理。

$$\boldsymbol{H}^t=\{\log_2(1+\frac{h_{i,j}^t}{h_{i,i}^t})\}_{i\in \boldsymbol{M},\,j\in \boldsymbol{M}} \tag{22}$$

因此在时隙t的状态设计为:$\boldsymbol{s}^t=\{\boldsymbol{H}^t,\boldsymbol{G}^t,r^{t-1},\boldsymbol{P}^{t-1}\}$。

动作:在式(8)中,功率是一个连续变量,只受最大功率约束。由于DQN的动作空间必须是离散的,因此将功率空间离散化为D个电平。除0 W的功率外,其他功率的最小值和最大值分别为:$P_{\min}$,$P_{\max}$。$P_{\min}$和$P_{\max}$单位为dBm。功率的动作空间离散化后如式(23)所示,包括0电平在内共有D种电平。链路i的发射端在时隙t的发射功率为p_i^t,$p_i^t\in \mathrm{P}^s$。为了更好地与式(6)相匹配和计算网络容量,P^s中的功率均转化以W为单位。所有链路在时隙t的发射功率为$P^t=\{p_i^t\}_{i\in \boldsymbol{M}}$。

$$\mathrm{P}^s=\left\{0,10^{\frac{P_{\min}}{10}-3},10^{\frac{P_{\min}}{10}+\frac{P_{\max}-P_{\min}}{10(D-1)}-3},\right.$$
$$10^{\frac{P_{\min}}{10}+\frac{2(P_{\max}-P_{\min})}{10(D-1)}-3},\cdots,10^{\frac{P_{\min}}{10}+\frac{(D-2)(P_{\max}-P_{\min})}{10(D-1)}-3},$$
$$\left.10^{\frac{P_{\min}}{10}+\frac{(D-1)(P_{\max}-P_{\min})}{10(D-1)}-3}\right\}\tag{23}$$

奖励：假定时隙t被预分配给链路j，在时隙t以每个链路的信道容量$\mathrm{C}^t=\{C_i^t\}_{i\in M}$作为基础设计奖励，全网奖励通过当前时刻的状态和根据公式(16)计算得到，其中S_m^t的计算方式如式(17)所示。通过奖励的设计保证优化问题中被分配链路的最小信道容量约束。根据式(16)我们可以认为奖励是在保证被分配链路可以有效传输数据的情况下，全网可以有效传输数据的总信道容量。如果不能保证被分配链路的最小信道容量，那么奖励只能是0。DQN在最大化Q值的过程中会保证被分配链路的最小信道容量。

在本节中，我们按照深度强化学习的框架设计了奖励、状态和动作来求解优化问题，确保了每个链路在被分配时隙的信息容量高于最小信道容量，并通过时隙复用和功率控制在每个时隙实现网络容量的最大化。由于DQN的奖励与链路及其分配的时隙相关，因此需要在每个时隙进行DQN的训练，即训练M个DQN。

在线执行阶段时，根据TDMA协议的时隙预分配时隙情况，直接将训练好的神经网络的参数$\boldsymbol{\theta}_i$加载到神经网络中运行。例如：在时隙1时将针对时隙1训练的神经网络参数加载到神经网络上。执行过程如图7所示。

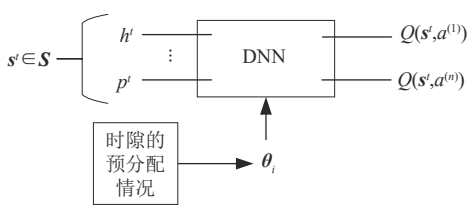


图7　串行DQN执行示意图

Fig. 7　Illustration of sequential DQN execution

3.3　基于DQN的时隙复用和功率控制协议

首先利用TDMA协议对每个链路分配一个时隙，然后在每个时隙利用设计的DQN控制所有网内链路功率以实现时隙复用和最大化全网信道容量。协议的流程如下所示：

(1)根据TDMA协议为每个链路分配一个时隙。

(2)根据3.2节设计的DQN奖励和状态动作空间训练每个时隙的DQN，DQN训练过程如表1所示。

表1　DQN训练流程

Tab. 1　DQN training process

DQN训练流程
初始化状态$\boldsymbol{s}_i^0\|i\in\boldsymbol{M}$，学习率$\alpha$，探索率最大值$\varepsilon_{\max}$，探索率最小值$\varepsilon_{\min}$
设置回放经验池$\boldsymbol{D}$和回放批次大小E
初始化M个神经网络参数为$\theta_i\|i\in\boldsymbol{M}$的动作-价值Q网络
初始化M个神经网络参数为$\theta_i^-=\theta_i$的目标Q网络
1.for episode=0,1⋯,I
2.　for t=0,1⋯,T
3.　　将$\boldsymbol{s}_i^t\|i\in\boldsymbol{M}$输入到目标Q网络得到不同动作的Q值集合$\{Q_{\boldsymbol{\pi}}(\boldsymbol{s}_i^t,\boldsymbol{a}_i;\boldsymbol{\theta}_i)\|\boldsymbol{a}\in P^s\}$；
4.　　根据自适应$\varepsilon-$greedy策略选择动作$\boldsymbol{a}_i^t$并更新ε[25]；
5.　　执行动作$\boldsymbol{a}_i^t$得到一个奖励r_i^t；
6.　　得到下一个时隙的状态$\boldsymbol{s}_i^{t+1}\|i\in\boldsymbol{M}$；
7.　　将$(\boldsymbol{s}_i^t,\boldsymbol{a}_i^t,r_i^t,\boldsymbol{s}_i^{t+1})$储存在经验回放池$\boldsymbol{D}_i$中
8.　　按照随机取样的方法取经验池$\boldsymbol{D}_i$中取出E批次经验$\boldsymbol{E}_{i\|i\in\boldsymbol{M}}$
9.　　for $\boldsymbol{E}$中每个经验组$(\boldsymbol{s}_i^j,\boldsymbol{a}_i^j,r_i^j,\boldsymbol{s}_i^{j+1})$
10.　　$y_i^j=r_i^j$
11.　　根据式(14)计算损失函数，根据式(15)更新θ_i
12.　　end for 经验回放结束
13.　　将θ_i^-更新为θ_i
14. end for
15. end for

(3)根据分配的时隙在线执行训练完成的DQN网络，例如在每帧的时隙1执行针对时隙1训练完成的DQN。

(4)将DQN执行的结果由地面站分发给各个无人机执行。

3.4　算法复杂度分析

在L_{layer}层DQN网络中，乘法的计算次数为：$N_{\text{multi}}=Ul_1+\sum_{j=1}^{L_{\text{layer}}-1}l_jl_{j+1}$，其中$U$是神经网络的输入层维度，$l_j$是第$j$层神经网络的神经元数。在训练阶段，有$I$个回合，每个回合包含$T$个时间步，训练的复杂度为$O(ITN_{\text{multi}})$。在执行阶段，每次执行策略网络的计算复杂度为$O(N_{\text{multi}})$。

4　仿真分析

4.1　仿真设置

在一个定向无人机网络中，我们假设有10个链路准备以固定的拓扑进行数据传输。10个链路的20个无人机随机部署在长宽为20 km×20 km，高度为1 km的区域内，并随机生成拓扑。每架无人机在部署的初始位置半径1 km内执行侦查任务，每进行50帧数据的传输随机移动到侦查范围的下一位置

继续侦查。利用Jakes模型来模拟动态的小尺度衰落信道，多普勒频率为2000 Hz。通过路径衰落公式模拟大尺度衰落[29]，工作频率为5 GHz。加性高斯白噪声为−105 dBm，发射功率最大为38 dBm，最小发射功率为5 dBm。功率等级D是10，为0~9，其中0为0功率，1为非零最小功率，9为最大功率。定向天线主瓣增益为20 dB，主瓣波束夹角为10°。10个链路收发端的定向天线均按照图3所示的方式指对方，且在通信过程中均按照拓扑指向对方。定向天线的传输距离是整个空间，当距离变大只会影响接收端的信干噪比，从而影响网络容量。每个链路的最小信道容量设置为全向通信在20 km×20 km×1 km范围内最大通信距离的五分之一时最大发射功率的无干扰信道容量。在本文中为：天线增益为0 dB，两个节点通信距离为通信距离5.67 km的信道容量，功率为38 dBm，干扰功率为0利用式(7)计算得到的信道容量。这么设置的原因是：全向通信距离较短因此将其最大通信距离设置为定向通信最大距离的五分之一，全向通信在最远距离无干扰的条件下进行最大功率通信才能保证正常业务传输所需的信道容量。本文利用Python实现环境的构建和算法的仿真。

在离线训练阶段，首先随机初始化DQN，然后前100轮进行随机探索。100轮后，根据自适应ε−greedy进入下一个探索期[28]。每轮训练有50个时隙，每10个时隙进行1次经验回放，经验回放随机抽取256个样本进行回放。本文采用Adam算法作为优化器，学习率从10^{-3}到10^{-4}衰减。DQN的所有参数设置如表2所示。

表2 DQN训练参数设置

Tab. 2 DQN training parameter configuration

参数	值	参数	值
每轮时隙数	50	初始学习率	0.001
观察轮数	100	最后学习率	0.0001
探索轮数	7900	初始探索$\varepsilon_{\max}$	0.3
训练时隙数	50	最终探索$\varepsilon_{\min}$	0.0001
记忆容量	50000	回放样本大小	256

根据TDMA协议，每个链路按照时隙号分配一个时隙，例如：链路1分配时隙1，链路2分配时隙2。在每个时隙执行设计的DQN算法和对比算法或协议。通过仿真了最大功率方法、随机功率方法、TDMA协议、三层FNN(Three-layer Feedforward Neural Network)的DQN和四层FNN(Four-layer Feedforward Neural Network)的DQN来评估基于DQN的定向无人机网络时隙复用和功率控制协议。三层FNN的DQN由一个大小为128×64的隐藏层、输入层、输出层组成。四层FNN的DQN的两个隐藏层大小为128×64和64×64。

(1)最大功率法(Maximal power)：每个链路在每个时隙均以最大功率进行发射；

(2)随机功率法(Random power)：每个链路在每个时隙按照随机功率进行发射；

(3)TDMA协议：每个链路分配一个时隙，在被分配的时隙链路按照最大功率发射，在其他时隙功率为0；

(4)Exclusive Region (ER)算法[30]：采用划定排他区域的方式实现多条链路在一个时隙并行通信；

(5)利用三层神经网络DQN执行3.3节提出的协议：基于三层神经网络的DQN在每个时隙进行功率控制；

(6)利用四层神经网络DQN执行3.3节提出的协议：基于四层神经网络的DQN在每个时隙进行功率控制。

4.2 一个时隙的链路复用和功率控制仿真分析

本节主要通过仿真展示一个时隙的链路复用情况和网络容量，以每帧第一个时隙为例进行仿真和分析。

图8是三层FNN的DQN网络和四层FNN的DQN的训练示意图，可以看到两个网络均到6000回合时在奖励为54时收敛，即：在时隙1保证链路1信道容量大于最小信道容量的情况下，全网能到达的有效信道容量在54左右。收敛后全网信道容量不断波动的原因主要是不断变化的信道状态和DQN仍在以很小的探索率在搜索最优解。

通过500帧的仿真，时隙1的全网有效信道容量示意图如图9所示，全网有效信道容量是满足式(8)约束条件的所有链路信道容量的和。图9可以看出在时隙1基于DQN的协议优于随机功率法、最大功率法和TDMA协议。最大功率法虽然增加了发射功率，但是由于各个节点间的互干扰增强导致信干噪比下降，因此性能差于提出的DQN协议。由于10个链路有10个功率等级，那么就有10^{10}种可能性，因此随机功率法找到最优解的概率极低。而TDMA协议虽然没有链路间的互干扰，但是没有利用定向网络多链路在同一时隙工作的特点，因此性能劣于提出的DQN协议。相较于TDMA协议，我们所设计的基于三层FNN的DQN的协议和基于四层FNN的DQN协议性能分别提高了239.01%和240.97%。

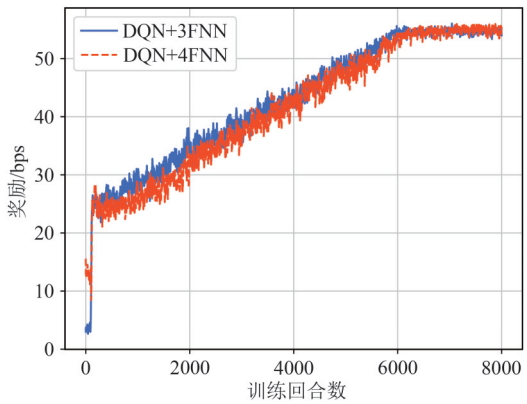


图 8　三层神经网络和四层神经网络的DQN在时隙1的训练情况

Fig. 8　Training illustration of three-layer and four-layer neural network-based DQNs in Slot 1

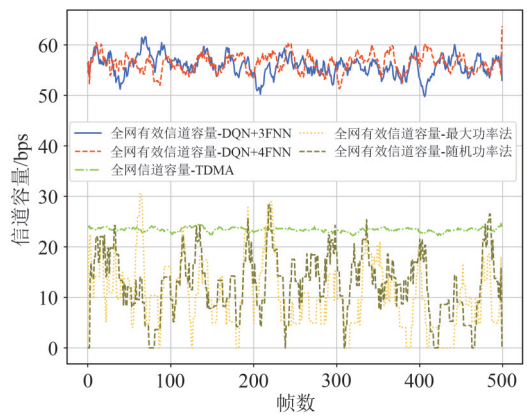


图 9　三层神经网络DQN协议、四层神经网络DQN协议和对比方法在每帧时隙1全网信道容量

Fig. 9　Network capacity comparison of the three-layer neural network DQN protocol, four-layer neural network DQN protocol, and comparative methods in slot 1 of each frame

为了展示所设计的DQN在一个时隙优化后的链路占用情况，在图10中给出了利用三层FNN的DQN和四层FNN的DQN进行优化后，所有链路在第一帧时隙1的功率等级。横坐标为链路序号，纵坐标为功率等级，0等级功率表示不占据该时隙。在图10中，在时隙1中通过三层FNN的DQN和四层FNN的DQN得到的每个链路发射功率不同，而在图9中可以看到三层FFN的DQN和四层FNN的DQN所得到的全网信道容量相近，因此虽然功率设置不同，但是这两种功率设置均是可以逼近最大全网有效信道容量的解。根据3.4节算法复杂度的分析，神经网络的层数越大，计算复杂度越高，因此在性能相近的情况下选择神经网络层数少的DQN以降低计算的复杂度。

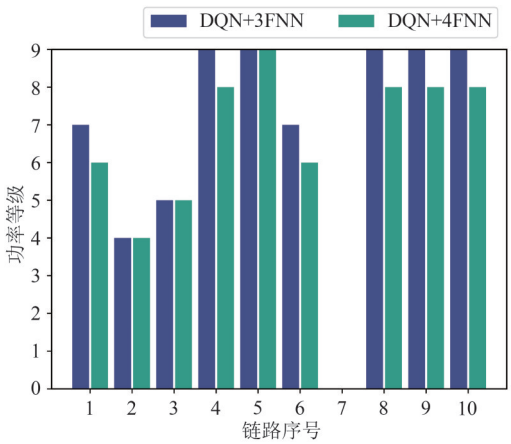


图 10　利用DQN优化后，每个链路在第一帧时隙1的功率等级

Fig. 10　Power levels of each link in slot 1 of the first frame after optimization using the DQN

4.3　一帧的链路复用和功率控制仿真分析

本节主要通过仿真对比方法和所提出的基于DQN的协议在一帧内全网信道容量以验证基于DQN协议的先进性。

我们通过进行5000个时隙的仿真，即500帧，计算对比方法和提出DQN协议的每帧的网络容量，仿真结果如图11所示。在图11中，ER+P9表示ER算法的功率等级为9。通过仿真，在我们设定的场景下，ER算法在功率等级为6时全网网络容量最大。相较于ER+P6，我们提出的基于三层FNN的DQN协议的网络容量提升了143.09%。

相较于TDMA协议，我们所设计的基于三层

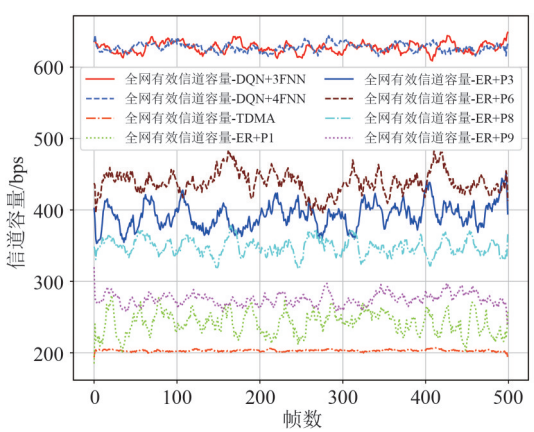
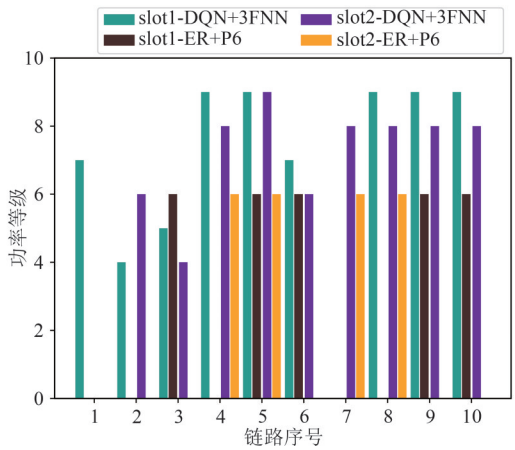


图 11　三层神经网络DQN协议、四层神经网络DQN协议和对比方法在每帧全网信道容量

Fig. 11　Network capacity comparison of the three-layer neural network DQN protocol, four-layer neural network DQN protocol, and comparative methods in each frame

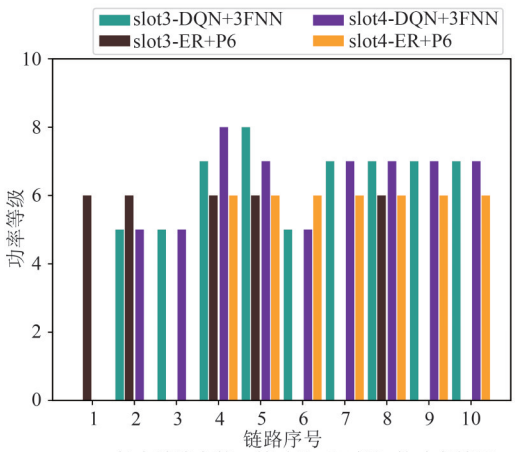
FNN的DQN的协议网络容量提高了308.70%。结合图9和图11,提出的基于DQN的时隙复用和功率控制协议可以大幅度提升定向无人机网络的网络容量。

为了展示在一帧内每个时隙每个链路的功率等级,以第一帧内10个时隙的例,图12是第一帧的所有链路利用DQN和ER+P6优化后所得到的功率



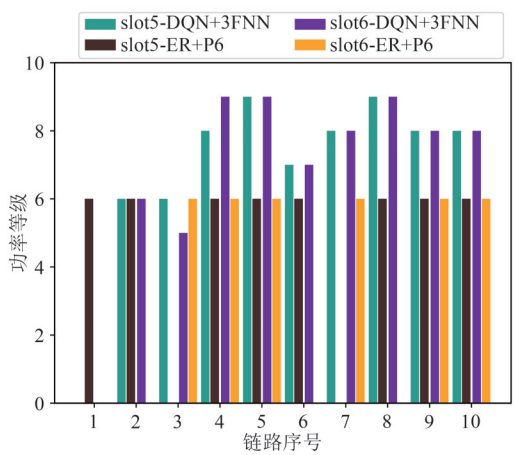
(a) 每个链路在第一帧时隙1和时隙2的功率等级

(a) Power levels of each link in slots 1 and 2 of the first frame



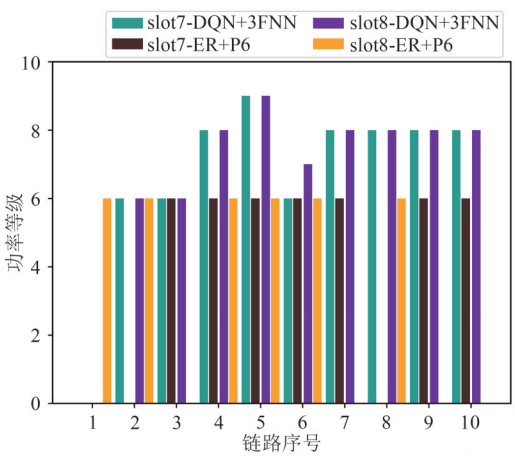
(b) 每个链路在第一帧时隙3和时隙4的功率等级

(b) Power levels of each link in slots 3 and 4 of the first frame



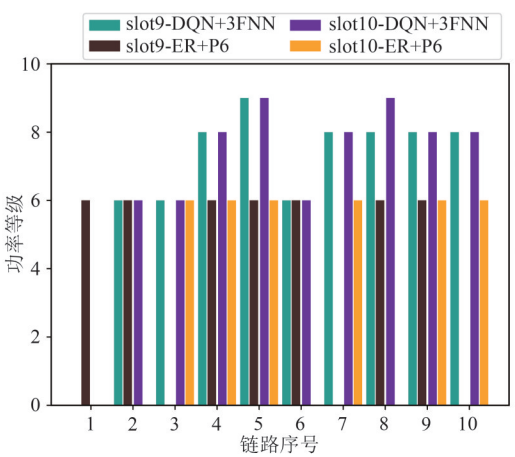
(c) 每个链路在第一帧时隙5和时隙6的功率等级

(c) Power levels of each link in slots 5 and 6 of the first frame



(d) 每个链路在第一帧时隙7和时隙8的功率等级

(d) Power levels of each link in slots 7 and 8 of the first frame



(e) 每个链路在第一帧时隙9和时隙10的功率等级

(e) Power levels of each link in slots 9 and 10 of the first frame

图12 第一帧每个时隙每个链路功率等级

Fig. 12 Power levels of each link in each slot of the first frame

等级。统计每个链路功率等级在第一帧不为0的时隙个数即在第一帧的时隙占用个数，所有链路在第一帧占用的时隙个数情况如图13所示。在图13中，两种方法均实现了每个链路在一帧内至少以最小的网络容量通信一次。然而，相较于ER+P6，本文提出的基于DQN的协议网络容量更大。在图12中，可以看到我们提出的基于DQN的协议可以根据时隙进行不同的功率优化，而ER协议只能选择时隙的通断以恒定的功率通信而不能根据信道环境动态的调整功率，这是在图11中基于DQN的协议优于ER协议的原因。

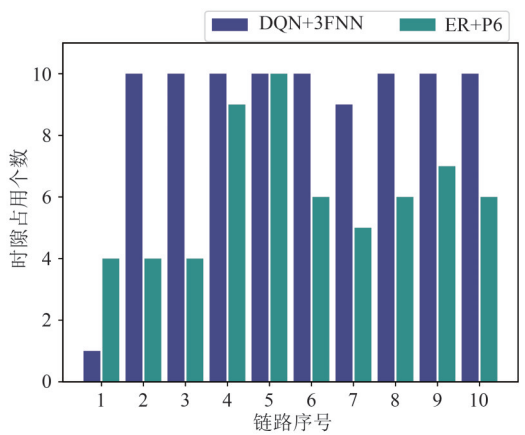


图13　第一帧每个链路占用时隙个数

Fig. 13　Number of time slots occupied by each link in the first frame

5　结论

定向天线为定向网络带来了时隙复用能力，这可以大大提升网络容量。然而，如何实现时隙复用是定向网络研究的难题。为此，本文提出了一种基于DQN的定向无人机网络时隙复用和功率控制协议。基于TDMA协议在每个时隙构建功率控制的优化问题，通过求解优化问题实现每个时隙的链路复用和功率控制以求最大化网络容量。为了利用DQN求解该优化问题，将优化问题的互干扰融入进DQN的状态设计，将网络容量融入奖励中以实现网络容量的最大化。仿真结果表明：相较于对比的协议，基于DQN的协议实现了更高的网络容量。

参考文献

[1] HU Jinqiang, WU Husheng, ZHAN Renjun, et al. Self-organized search-attack mission planning for UAV swarm based on wolf pack hunting behavior[J]. Journal of Systems Engineering and Electronics, 2021, 32(6): 1463-1476.

[2] XIN Hongbo, CHEN Qingyang, WANG Yujie, et al. A path planning and guidance method for multi-UAVs coordinated strike with time-space constraints[C]//2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). Hangzhou, China. IEEE, 2020: 193-198.

[3] MAHMUD M T, RAHMAN M O, ALQAHTANI S A, et al. Cooperation-based adaptive and reliable MAC design for multichannel directional wireless IoT networks [J]. IEEE Access, 2021, 9: 97518-97538.

[4] COYLE A. Using directional antenna in UAVs to enhance tactical communications[C]//2018 Military Communications and Information Systems Conference (MilCIS). Canberra, ACT, Australia. IEEE, 2018: 1-6.

[5] SAMANDARI A, WILLIG A. TDMA slot allocation for UAV formations: minimum superframe lengths for two-dimensional equidistant deployments[C]//2022 32nd International Telecommunication Networks and Applications Conference (ITNAC). Wellington, New Zealand. IEEE, 2023: 56-63.

[6] QIAO Jian, CAI Lin X, SHEN Xuemin, et al. STDMA-based scheduling algorithm for concurrent transmissions in directional millimeter wave networks [C]//2012 IEEE International Conference on Communications (ICC). Ottawa, ON, Canada. IEEE, 2012: 5221-5225.

[7] RAKESH R T, DAS G, SEN D. Energy efficient scheduling for concurrent transmission in millimeter wave WPANs [J]. IEEE Transactions on Mobile Computing, 2018, 17(12): 2789-2803.

[8] 谢添，高士顺，赵海涛，等. 基于强化学习的定向无线通信网络抗干扰资源调度算法[J]. 电波科学学报，2020,35(4):531-541.
XIE Tian, GAO Shishun, ZHAO Haitao, et al. An anti-jamming resource scheduling algorithm for directional wireless communication networks based on reinforcement learning [J]. Chinese Journal of Radio Science, 2020,35(4):531-541. (in Chinese)

[9] WANG Haijun, JIANG Bo, ZHAO Haitao, et al. Joint resource allocation on slot, space and power towards concurrent transmissions in UAV ad hoc networks [J]. IEEE Transactions on Wireless Communications, 2022, 21(10): 8698-8712.

[10] IBRAHIM A, ALFA A S. Using Lagrangian relaxation for radio resource allocation in high altitude platforms

[J]. IEEE Transactions on Wireless Communications, 2015, 14(10): 5823-5835.

[11] ZHANG Shuo, SHI Shuo, GU Shushi, et al. Power control and trajectory planning based interference management for UAV-assisted wireless sensor networks[J]. IEEE Access, 2019, 8: 3453-3464.

[12] LIU Chang, QIN Xiaowei, ZHANG Sihai, et al. Proportional-fair downlink resource allocation in OFDMA-based relay networks[J]. Journal of Communications and Networks, 2011, 13(6): 633-638.

[13] HO W W L, LIANG Yingchang. Optimal resource allocation for multiuser MIMO-OFDM systems with user rate constraints[J]. IEEE Transactions on Vehicular Technology, 2009, 58(3): 1190-1203.

[14] GORTZEN S, SCHMEINK A. Non-asymptotic bounds on the performance of dual methods for resource allocation problems[J]. IEEE Transactions on Wireless Communications, 2014, 13(6): 3430-3441.

[15] QIAN Li ping, WU Yuan, JI Bo, et al. Optimal ADMM-based spectrum and power allocation for heterogeneous small-cell networks with hybrid energy supplies[J]. IEEE Transactions on Mobile Computing, 2021, 20(2): 662-677.

[16] 潘筱茜, 张姣, 刘琰, 等. 基于深度强化学习的多域联合干扰规避[J]. 信号处理, 2022, 38(12): 2572-2581.
PAN Xiaoqian, ZHANG Jiao, LIU Yan, et al. Multi-domain joint interference avoidance based on deep reinforcement learning[J]. Journal of Signal Processing, 2022, 38(12): 2572-2581. (in Chinese)

[17] YU Kaiwen, ZHAO Chonghao, WU Gang, et al. Distributed two-tier DRL framework for cell-free network: association, beamforming and power allocation[EB/OL]. 2023: arXiv: 2303.12479. https://arxiv.org/abs/2303.12479.pdf.

[18] CHALLITA U, SAAD W, BETTSTETTER C. Interference management for cellular-connected UAVs: A deep reinforcement learning approach[J]. IEEE Transactions on Wireless Communications, 2019, 18(4): 2125-2140.

[19] LI Lixin, CHENG Qianqian, XUE Kaiyuan, et al. Downlink transmit power control in ultra-dense UAV network based on mean field game and deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2020, 69(12): 15594-15605.

[20] NASIR Y S, GUO Dongning. Deep reinforcement learning for joint spectrum and power allocation in cellular networks[C]//2021 IEEE Globecom Workshops (GC Wkshps). Madrid, Spain. IEEE, 2022: 1-6.

[21] XU Wenjun, LEI Huangchun, SHANG Jin. Joint topology construction and power adjustment for UAV networks: A deep reinforcement learning based approach[J]. China Communications, 2021, 18(7): 265-283.

[22] ZHANG Chiya, LIANG Shiyuan, HE Chunlong, et al. Multi-UAV trajectory design and power control based on deep reinforcement learning[J]. Journal of Communications and Information Networks, 2022, 7(2): 192-201.

[23] 赵国锋, 卢奕杉, 徐川, 等. 面向航天器有线无线混合场景的流调度机制研究[J]. 电子与信息学报, 2023, 45(2): 464-471.
ZHAO Guofeng, LU Yishan, XU Chuan, et al. Research on flow scheduling mechanism for spacecraft wired wireless hybrid scenario[J]. Journal of Electronics & Information Technology, 2023, 45(2): 464-471. (in Chinese)

[24] 胡静. 基于深度强化学习的多小区OFDMA系统资源分配方法研究[D]. 南京: 南京邮电大学, 2022.
HU Jing. Research on deep reinforcement learning based resource allocation for multi-cell OFDMA systems[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2022. (in Chinese)

[25] YE Hao, LI G Y, JUANG B H F. Deep reinforcement learning based resource allocation for V2V communications[J]. IEEE Transactions on Vehicular Technology, 2019, 68(4): 3163-3173.

[26] WU Qingqing, ZENG Yong, ZHANG Rui. Joint trajectory and communication design for multi-UAV enabled wireless networks[J]. IEEE Transactions on Wireless Communications, 2018, 17(3): 2109-2121.

[27] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

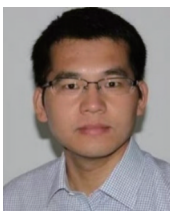
[28] SUTTON R S, BARTO A G. Reinforcement Learning: An Introduction[M]. 2nd ed. Cambridge, MA: MIT Press, 2018.

[29] 徐文染, 陈燚涛. 基于Friis传输公式的RSSI测距模型研究[J]. 武汉纺织大学学报, 2022, 35(4): 38-42.
XU Wenran, CHEN Yitao. Research on RSSI ranging model based on friis transmission formula[J]. Journal of Wuhan Textile University, 2022, 35(4): 38-42. (in Chinese)

[30] CAI L X, CAI Lin, SHEN Xuemin, et al. Rex: A randomized exclusive region based scheduling scheme for mmWave WPANs with directional antenna[J]. IEEE Transactions on Wireless Communications, 2010, 9(1): 113-121.

作者简介

梁仕杰　男，1995年生，河北石家庄人。中国电子科技集团公司第五十四研究所博士研究生，主要研究方向为认知无线电网络和定向组网。
E-mail：shijieliang21@163.com



赵海涛　男，1981年生，山东昌乐人。国防科技大学电子科学学院教授，主要研究方向为认知无线电网络和自组织网络。
E-mail：haitaozhao@nudt.edu.cn

张　姣　女，1990年生，湖南汨罗人。国防科技大学电子科学学院讲师，主要研究方向为边缘计算和无人机组网。
E-mail：zhangjiao16@nudt.edu.cn

王海军　男，1993年生，安徽淮北人。国防科技大学电子科学学院讲师，主要研究方向为无人机通信和组网。
E-mail：haijunwang14@nudt.edu.cn

魏急波　男，1968年生，湖北汉川人。国防科技大学电子科学学院教授，主要研究方向为软件无线电和认知通信网络。
E-mail：wjbhw@nudt.edu.cn

王俊芳　男，1963年生，河北盐山人。中国电子科技集团公司第五十四研究所研究员，主要研究方向为综合电子信息系统和通信网络理论。
E-mail：jfwang63@163.com

（责任编辑：边熙淳）